Report on K-Nearest Neighbors and Hierarchical Clustering

Introduction

This report presents the results of clustering analyses performed on two datasets: the Iris dataset and the Wine dataset. The clustering techniques employed include K-Nearest Neighbors (KNN) and Hierarchical Clustering using Ward's method. The performance of the clustering algorithms was evaluated using the Silhouette Score and the Davies-Bouldin Score, metrics that measure the quality of the clusters formed.

Iris Dataset Analysis

Dataset Overview

The Iris dataset consists of 150 samples from three species of Iris flowers (Iris setosa, Iris versicolor, and Iris virginica), with four features measured for each sample: sepal length, sepal width, petal length, and petal width.

Clustering Techniques

K-Means was applied to cluster the Iris dataset, with different distance metrics evaluated. The following results were obtained:

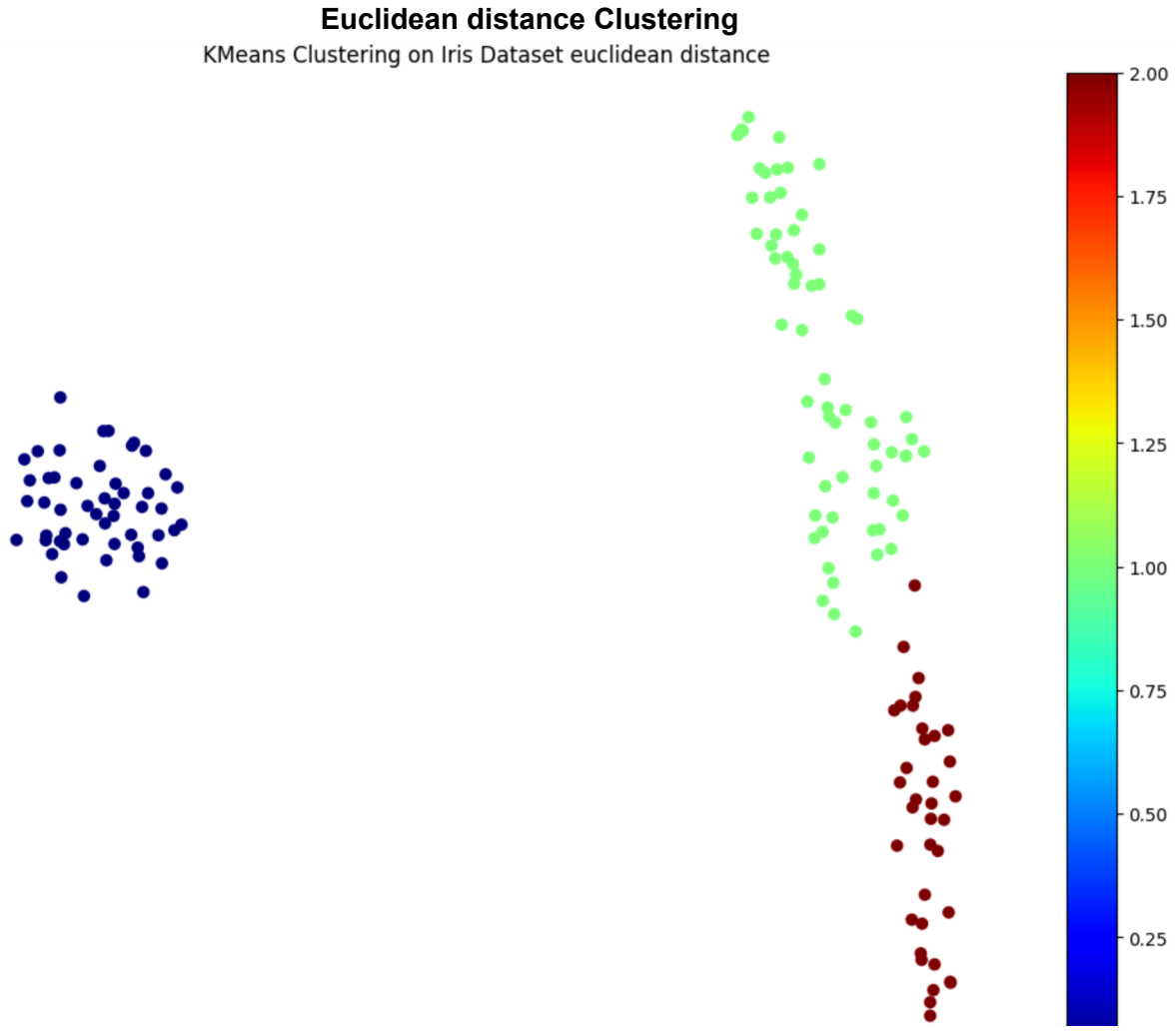
- Results for Euclidean Distance:
  - Silhouette Score: 0.5202
  - Davies-Bouldin Score: 0.6686
- Results for Manhattan Distance:
  - Silhouette Score: 0.5296
  - Davies-Bouldin Score: 0.6762
- Results for Cosine Distance:
  - Silhouette Score: 0.7471
  - Davies-Bouldin Score: 0.7775

The highest Silhouette Score was achieved using Cosine Distance, indicating better-defined clusters compared to the other distance metrics. Additionally, Hierarchical Clustering using Ward's method was applied to the Iris dataset.

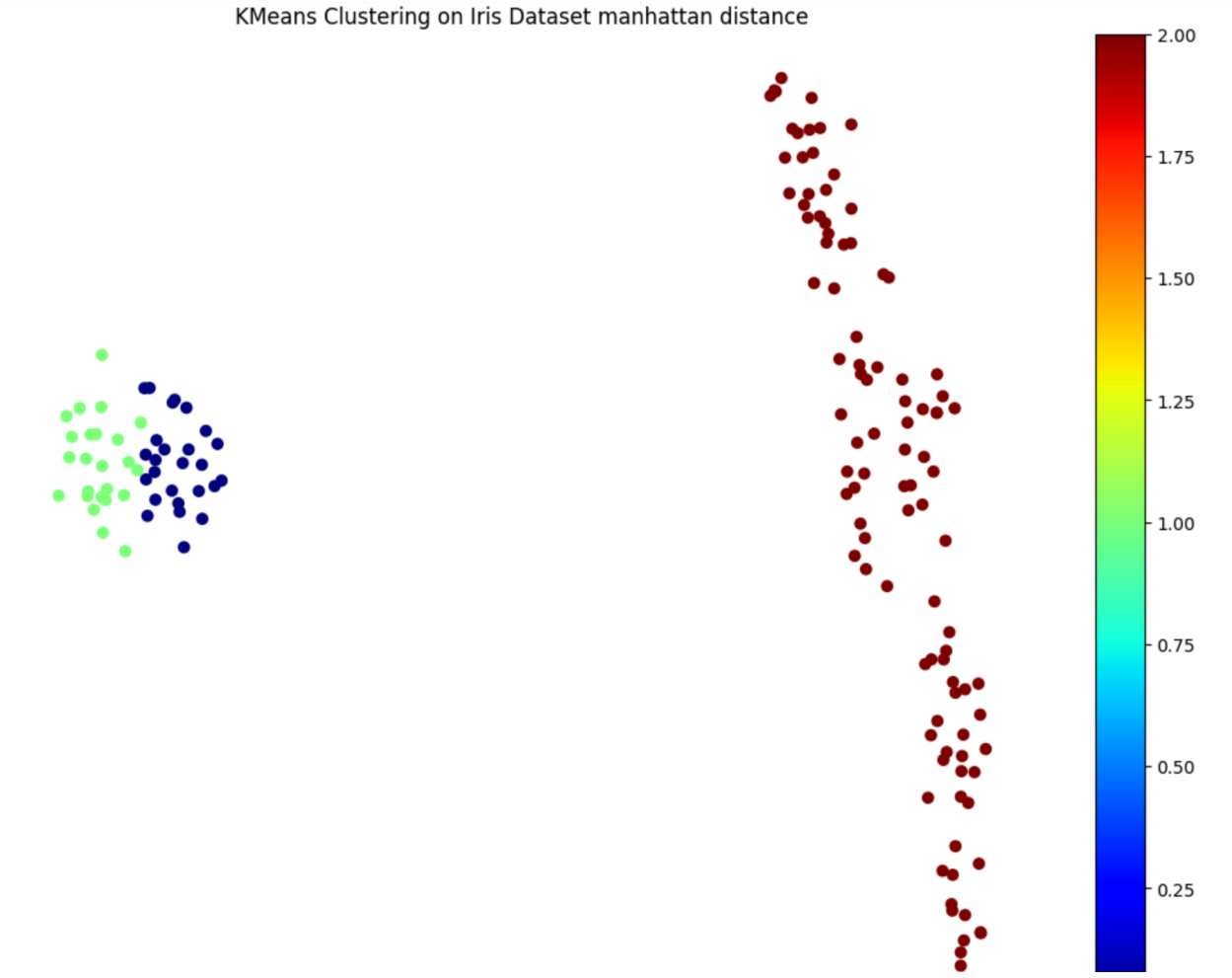
- Silhouette Score for Hierarchical Clustering: 0.56

This score suggests that the clusters formed using hierarchical clustering were relatively distinct, providing good separation between the species.

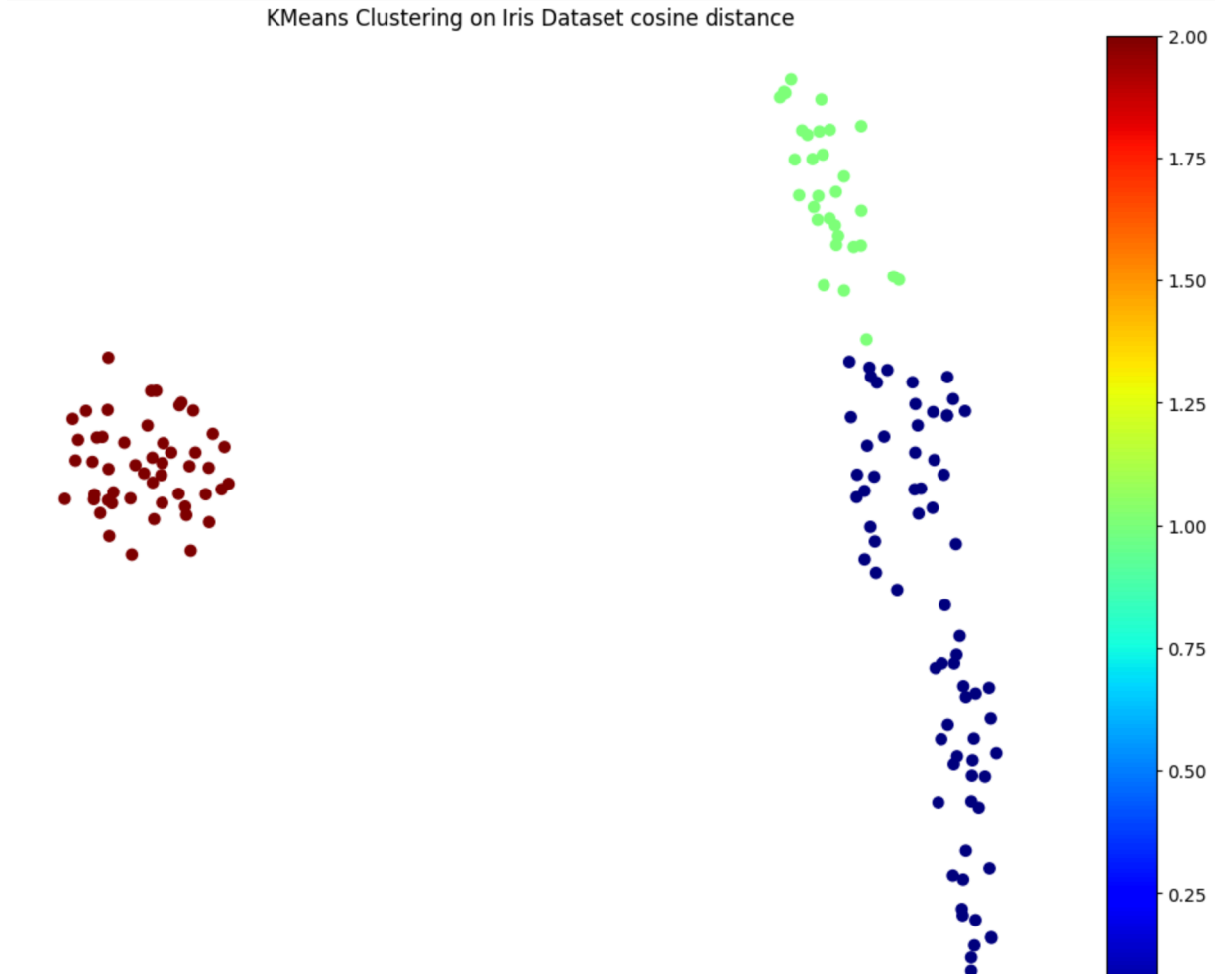
Clustering Graphs and Dendrograms



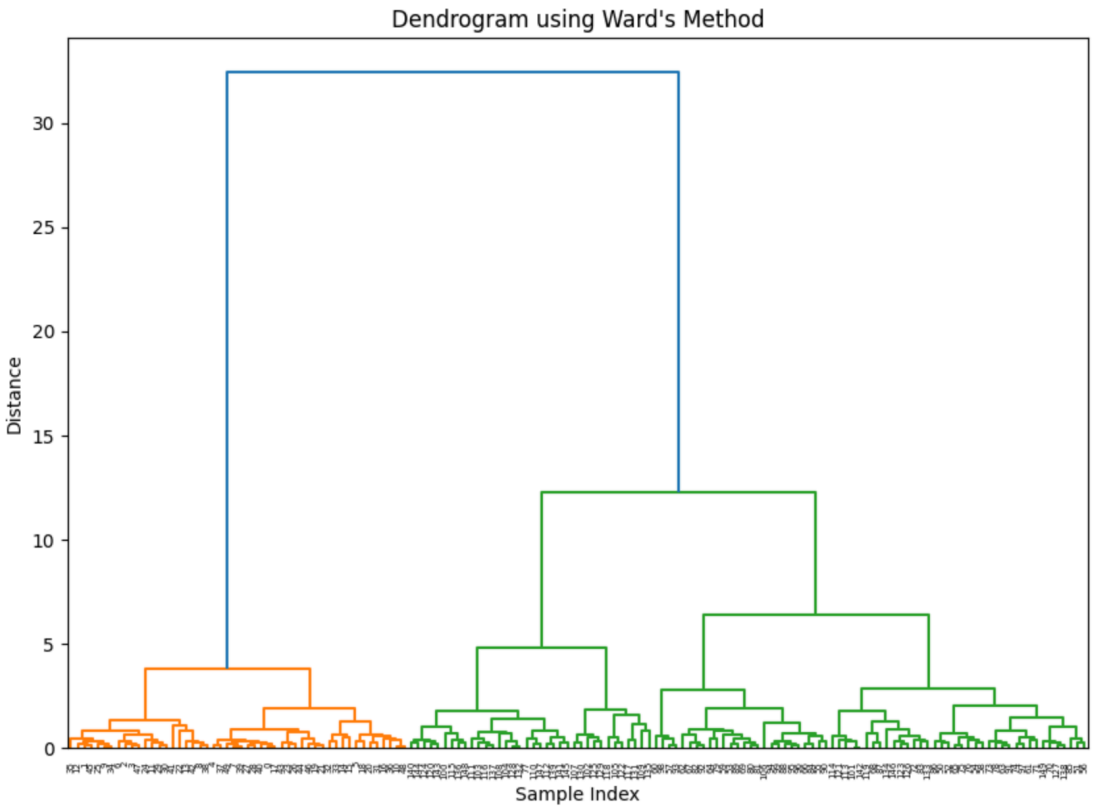
Manhattan distance Clustering



Cosine similarity Clustering



Hierarchical Clustering Dendrogram



Wine Dataset Analysis

Dataset Overview

The Wine dataset contains 178 samples with 13 features describing various chemical properties of wines. The dataset is used to classify wines into different cultivars based on these attributes.

Data Preprocessing

No normalization was applied to the Wine dataset before clustering.

Clustering Techniques

Similar to the Iris dataset, K-Means was used with various distance metrics:

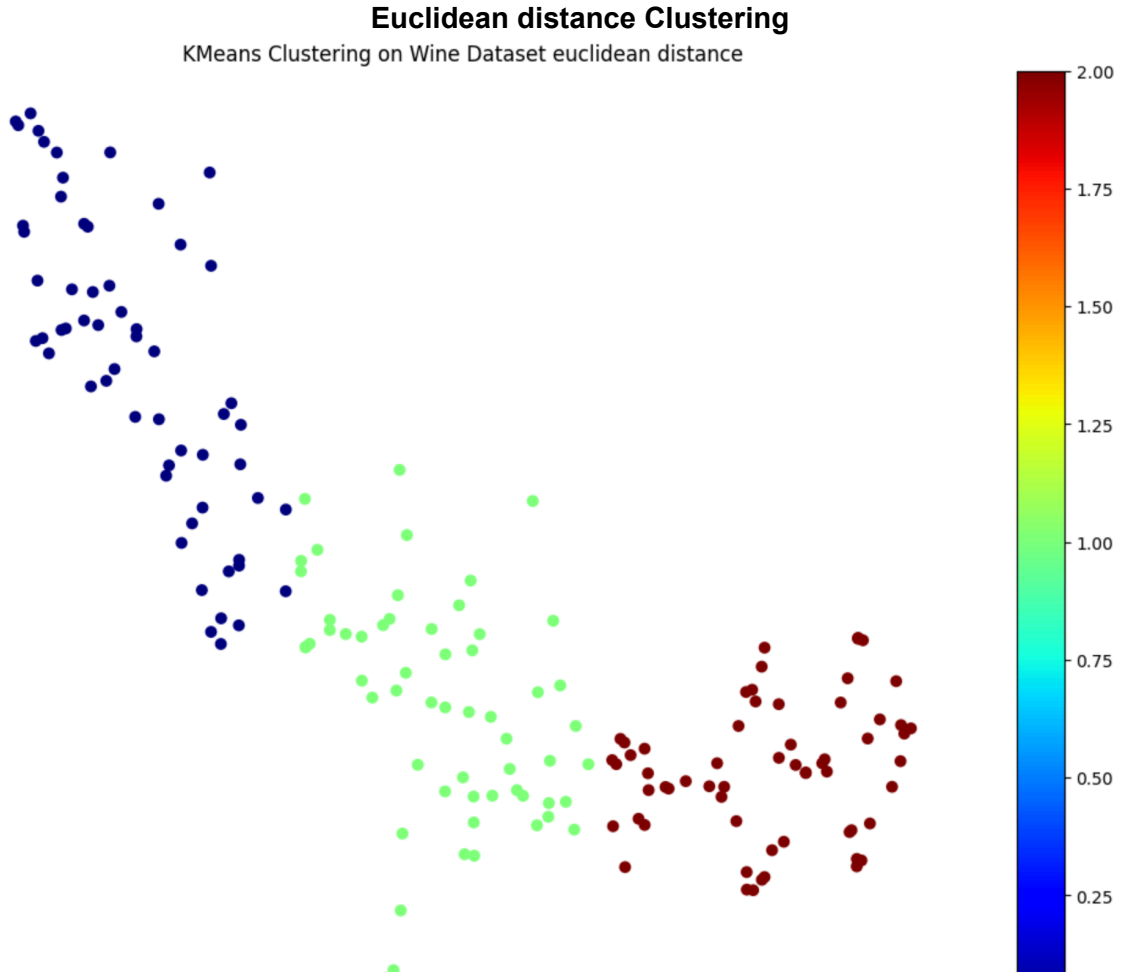
- Results for Euclidean Distance:
  - Silhouette Score: 0.5666
  - Davies-Bouldin Score: 0.5292
- Results for Manhattan Distance:
  - Silhouette Score: 0.5268
  - Davies-Bouldin Score: 0.5301
- Results for Cosine Distance:
  - Silhouette Score: 0.6960
  - Davies-Bouldin Score: 0.6643

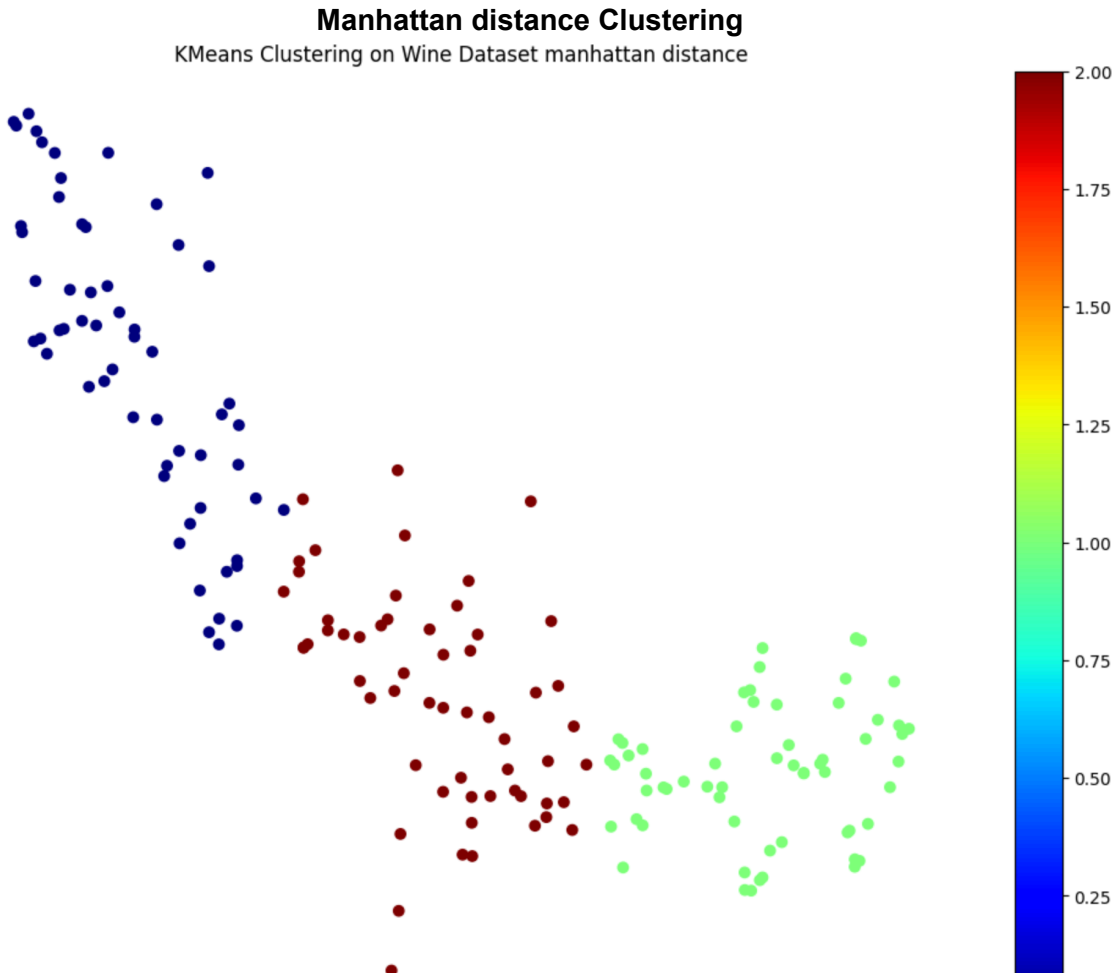
The Silhouette Scores for the Wine dataset indicate that the clusters are distinct, particularly with the Euclidean Distance yielding the highest score among the metrics evaluated. Hierarchical Clustering using Ward's method was also applied to the Wine dataset:

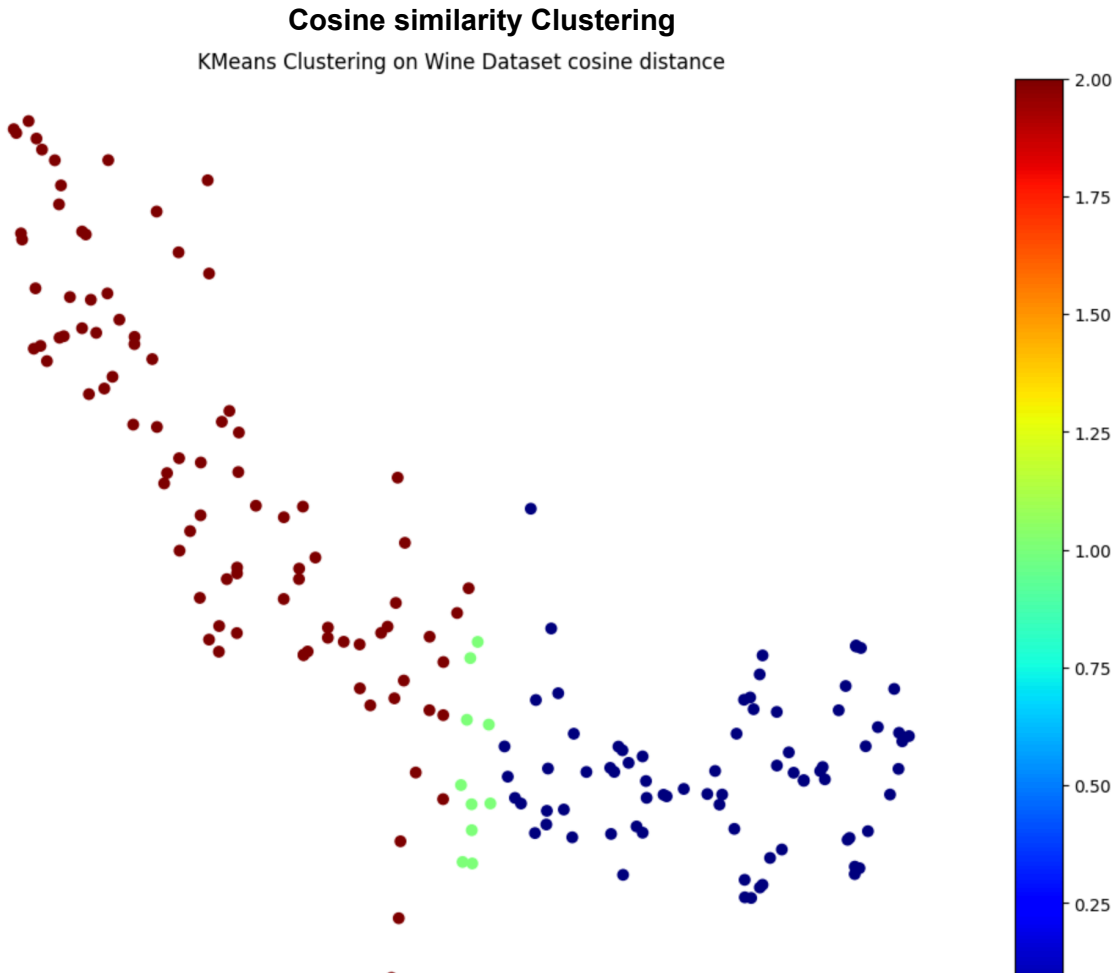
- Silhouette Score for Hierarchical Clustering: 0.56

This score reflects that the hierarchical clustering produced well-defined clusters for the Wine dataset, indicating good separation.

Clustering Graphs and Dendrograms




Manhattan distance Clustering
KMeans Clustering on Wine Dataset manhattan distance
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25


Cosine similarity Clustering
KMeans Clustering on Wine Dataset cosine distance
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25

Hierarchical Clustering Dendrogram.